Integrating virtualisation technologies within the CERN IT-FIO fabric

HEPiX 2009 Umeå

Ewan Roche

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Overview

- Quattor
  - What it is
  - How it models virtual machines
  - Using “Alien” Technology
- Virtualising Servers
  - An example “problem” cluster
  - Current issues
- Using virtualisation in batch systems
  - Current issues
  - The way forwards

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

What is Quattor?

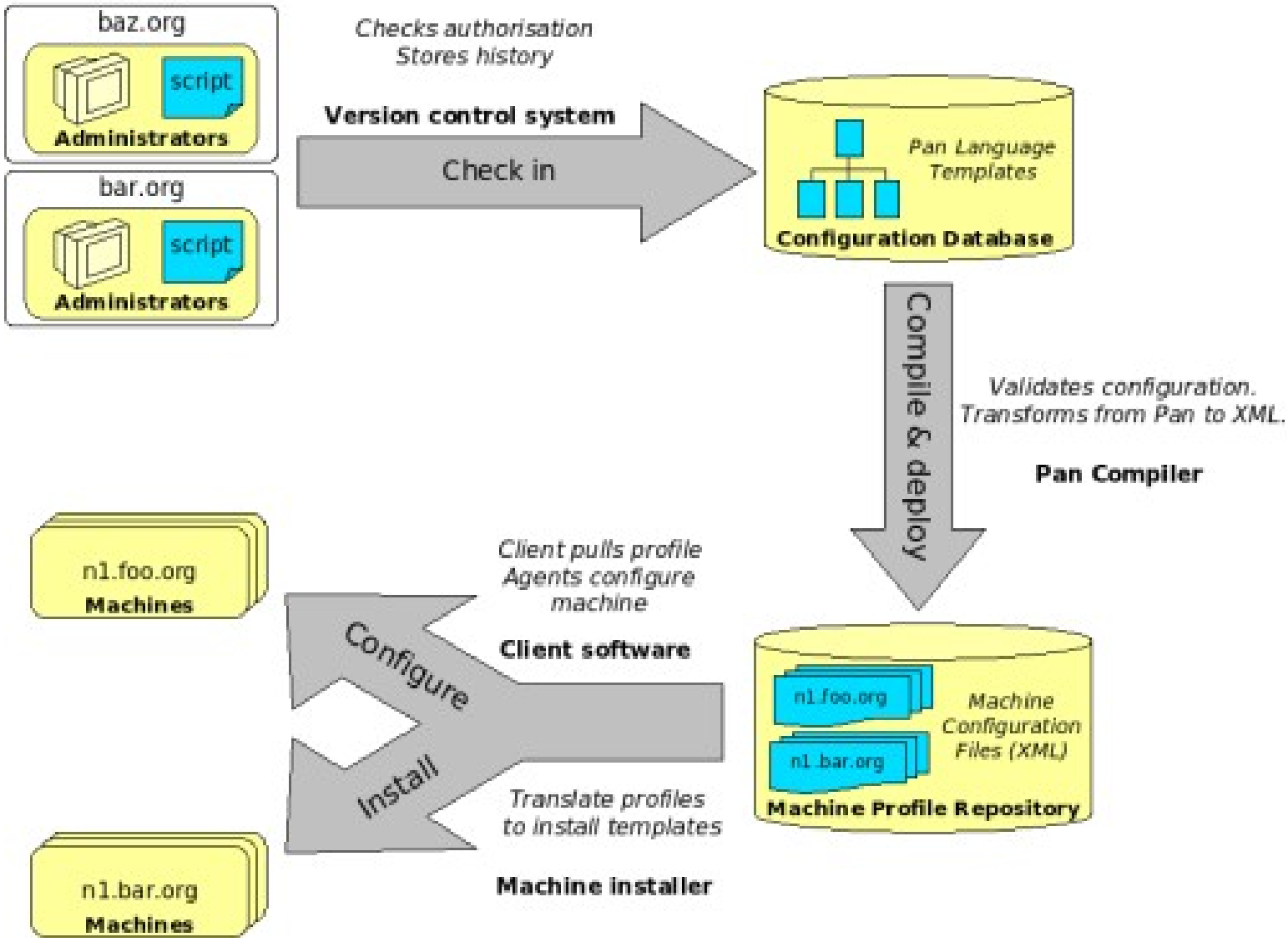
In Quattor, templates describing computers and services are written in a declarative language called Pan.

The set of templates describing the configuration of a complete system are usually stored in a configuration database that provides user authorisation and version control.

Pan templates are compiled to XML profiles which are then deployed to managed machines.

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

What is Quattor?



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Full Integration in Quattor I

The relationship between guests and hosts is defined in the host template

```
variable XEN_GUESTS_NAME = list("lxvm1000","lxvm1003");

"/system/enclosure/children" = XEN_GUESTS_NAME;

"/system" = block_add_list(XEN_GUESTS_NAME, XEN_VG);

"/software/components" = vm_add_list(XEN_GUESTS_NAME, XEN_VG);
```

Functions then act on the generated XML to create the relevant logical volumes and guest configuration files – Xen only at present.

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Full Integration in Quattor II

The same model as for enclosures (i.e blades in a rack) has been used to describe virtual machines

With the “Full Integration” approach and local disk we easily reach a situation where we have 12+ guests in a single “enclosure”

Conceptually this is fine but blades don't move very often but we hope virtual machines will.....

Any maintenance on this enclosure requires the shutdown of 12+ guests (unless we can migrate them)

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Using “Alien” Technology I

Virtualisation offers a number of desirable features

- Live Migration
- Load Balancing
- High Availabilty

These rely on shared storage and are, at present, only available in commercial products

We wish to keep using Quattor to manage “machines”

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Using “Alien” Technology II

How can we use these products whilst avoiding vendor lock-in and without too much work?

The best approach seems to be to hide the virtualisation from Quattor so it simply does what it's good at. Quattor will not know if something is a real or virtual machine.

Integration with monitoring and some tools required but this should be straightforward

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Using “Alien” Technology III

The majority of VM solutions rely on the concept of the machine image – this is entirely different to the individual machine profile view of Quattor.

As such we need to create a VM image that PXE boots to begin the usual build process

Alternatives include the VMWare virtual BIOS which supports PXE booting

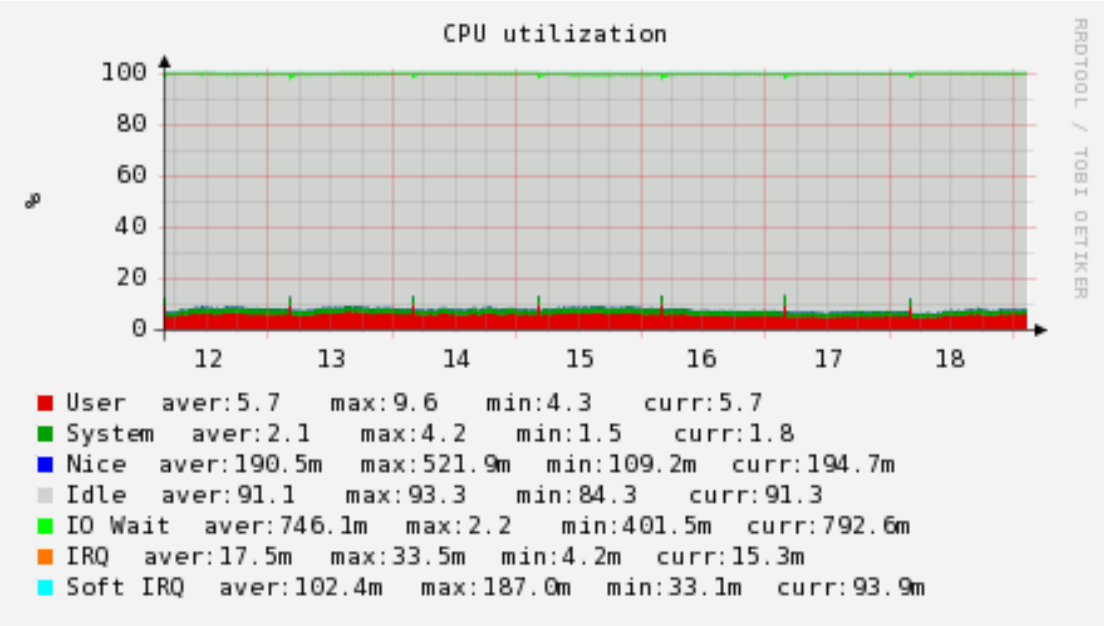
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

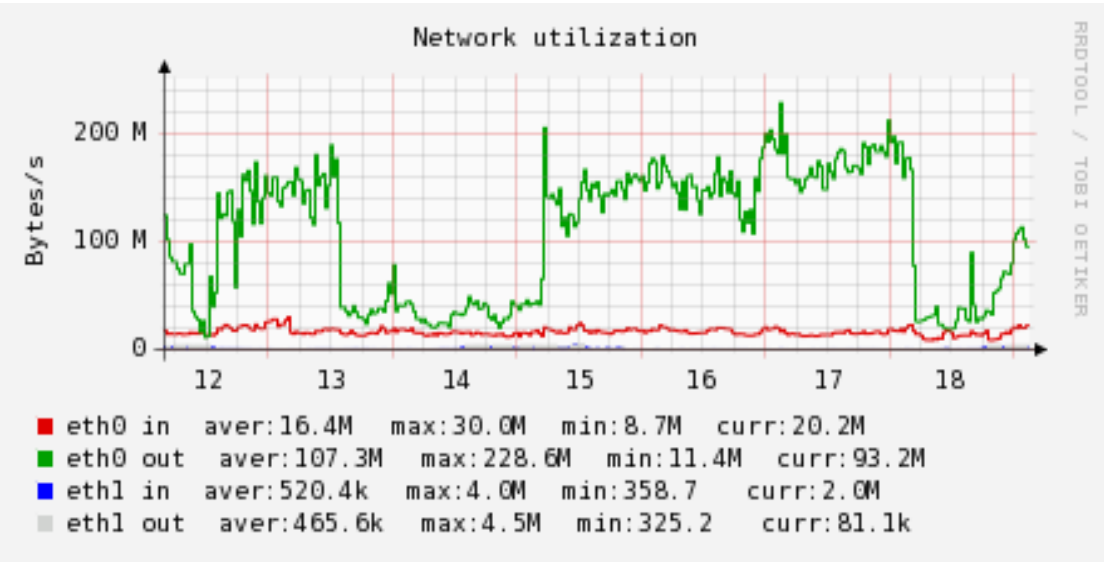
Overview

- Quattor
  - What it is
  - How it models virtual machines
  - Using “Alien” Technology
- Virtualising Servers
  - An example “problem” cluster
  - Current issues
- Using virtualisation in batch systems
  - Current issues
  - The way forwards

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Consolidation I





VOBox Cluster

156 hosts

Stateful services

Downtime is a problem

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Consolidation II

A VOBox is a dedicated machine managed in collaboration between ourselves and the experiment

These run important experiment specific services such as job submission and tracking

Scheduling downtime (e.g. for hardware maintenance) is difficult for both sides

Resource allocation is another issue especially when legacy operating systems are needed

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Consolidation III

Virtualisation offers the chance to break out of the retirement cycle dictated by hardware

Batches of hardware reach the end of warranty and need to be removed – currently this is painful for such dedicated machines

Such retirements could be very difficult during LHC running.

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Consolidation IV

Virtual Machines will be created and added into CERN's existing hardware management system (HMS). From this point on we will have total equivalence between real and virtual machines.

Any magic, such as load balancing, is hidden from Quattor.

We are still considering the options for the shared storage that underpins many features.

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Consolidation V

Reliability is a major consideration. The entire cluster failing is not an option

- For the VOBox cluster the current single points of failure are
  - Core routers
  - Computer Centre UPS
- We will need to provide a comparable level of reliability
  - Shared storage
  - Machine orchestration

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Overview

- Quattor
  - What it is
  - How it models virtual machines
  - Using “Alien” Technology
- Virtualising Servers
  - An example “problem” cluster
  - Current issues
- Using virtualisation in batch systems
  - Current issues
  - The way forwards

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Virtualising Batch I

We face a number of problems with our batch farm

- Maintenance
  - Reboots for new kernels / core libraries
  - Do not wish to apply updates to running jobs
- Operating system ecosystem
  - New hardware requires SL5
  - Workload still needs SL4
  - The required SL4/SL5 ratio changes
- Protecting users from eachother
  - A job sees 8 cores so it will try and use them
  - Memory limits
  - Crashed jobs can leave orphan/ghost processes

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Virtualising Batch II

The consolidation approach is fine for service machines but is not well suited to use on the batch farm.

There are large advantages to frequently rebuilding batch nodes (per job or per day)

Rebuilding each machine once per day would cause a large drop in efficiency – once per job is even worse (especially for pilot jobs)

This is where an image based approach is good but how do we create and maintain the images?

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Virtualising Batch III

We will have some traditionally installed guests configured as batch nodes that are frequently updated

From these we will create template images once per day

Launched images will start to drain after 24 hours and once running jobs have finished they will shutdown

As such the VM lifetime is between 24 hours and 2 weeks

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Virtualising Batch IV

The queue composition is monitored and the composition of the cluster changed accordingly

There is nothing to stop us running some part of batch in this manner and the rest in a more traditional way if desired

If needed we can work with experiments to provide custom images that we control

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Current Projects I

We are evaluating Platform VMO to provide the intelligence for VM placement, migration and load balancing.

Server consolidation - Citrix XenServer

- Good Performance
- Live Migration
- But it is a new OS to maintain

Batch - RedHat/SL Xen (eventually KVM)

- Already supported
- No need for migration

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Summary

- Current virtualisation scheme is not sustainable
  - Need to embrace migration and other tools
- Potentially huge benefits in ease of management
  - Batch becomes easy to “reboot”
  - Downtime for HW interventions ceases
  - HW retirement become much less of an issue
- Clear problems have been identified for which virtualisation offers a nice solution
- This is development work – see the next HEPiX for an update

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

THE END

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it